International Journal of Computer Science Engineering
and Information Technology Research (IJCSEITR)
ISSN(P): 2249-6831; ISSN(E): 2249-7943
Vol. 4, Issue 3, Jun 2014, 99-102
© TJPRC Pvt. Ltd.

TRANS STELLAR
• Journal Publications • Research Consultancy

AUTOMATIC METADATA HARVESTING FROM DIGITAL CONTENT

RUSHABH D. DOSHI[1] & GIRISH H MULCHANDANI[2]

[1]M.E Student, Department of Computer Engineering, V.V.P Engineering College, Rajkot, Gujarat, India

[2]Lecturer, V.V.P Engineering College, Rajkot, Gujarat, India

ABSTRACT

Metadata Extraction s one of the predominant research fields in information retrieval. Metadata is used to references information resources. Most metadata extraction systems are still human intensive since they require expert decision to recognize relevant metadata but this is time consuming. However automatic metadata extraction techniques are developed but mostly works with structured format. We proposed a new approach to harvesting metadata from document using NLP. As NLP stands for Natural Language Processing work on natural language that human used in day today life.

KEYWORDS: Metadata, Extraction, NLP, Grammars

INTRODUCTION

Metadata is data that describes another data Metadata describes an information resource, or helps provide access to an information resource. A collection of such metadata elements may describe one or many information resources. For example, a library catalogue record is a collection of metadata elements, linked to the book or other item in the library collection through the call number. Information stored in the "META" field of an HTML Web page is metadata, associated with the information resource by being embedded within it. The key purpose of metadata is to facilitate and improve the retrieval of information. At library, college, Metadata can be used to achieve this by identifying the different characteristics of the information resource: the author, subject, title, publisher and so on. Various metadata harvesting techniques is developed to extract the data from digital libraries.

NLP is a field of computer science, artificial intelligence and linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human computer interaction. Recent research has increasingly focused on unsupervised and semi-supervised learning algorithms. Such algorithms are able to learn from data that has not been hand-annotated with the desired answers, or using a combination of annotated and non-annotated data. The goal of NLP evaluation is to measure one or more qualities of an algorithm or a system, in order to determine whether (or to what extent) the system answers the goals of its designers, or meets the needs of its users.

METHOD

In this paper we proposed automatic metadata harvesting algorithm using natural language (i.e. humans used in day today works). Our technique is rule based. So it does not require any training dataset for it. We harvest metadata based on English Grammar Terms. We identify the possible set of metadata then calculate their frequency then applying weight term based on their position or format that apply to it. The rest of the paper is organized as follows. The next section review some related work regarding to metadata harvesting from digital content. Section gives the detailed description of proposed idea presented here. At last paper is concluded with summary.

RELATED WORK

Existing Metadata harvesting techniques are either machine learning method or ruled based methods. In machine learning method set of predefined template that contains dataset are given to machine to train machine. Then machine is used to harvest metadata from document based on that dataset. While in rule based method most of techniques set ruled that are used to harvest metadata from documents.

In machine learning approach extracted keywords are given to the machine from training documents to learn specific models then that model are applied to new documents to extract keyword from them. Many techniques used machine learning approach such as automatic document metadata extraction using support vector machine.

In rule based techniques some predefined rules are given to machine based on that machine harvest metadata from documents. Positions of word in document, specific keyword are used as category of document and etc. are examples rules that are set in various metadata harvest techniques. In some case Metadata classification is based on document types (e.g. purchase order, sales report etc.) and data context (e.g. customer name, order date etc.) [1].

Other statistical methods include word frequency [2], TF*IDF [3], word co-occurrences [4]. Later on some techniques are used to harvest key phrase based on TF*PDF [5]. Other techniques use TDT (Topic Detection and Tracking) with aging theory to harvest metadata from news website [6]. Some techniques used DDC/RDF editor to define and harvest metadata from document and validate by thirds parties [7]. Several models are developed to harvest metadata from corpus. Now days most of techniques used models that all are depends on corpus.

Proposed Theory

Our approach focused on harvesting a metadata from document based on English grammar. English grammar has many categories which categorized the word in statement. Grammar categories such as NOUN,VERB, ADJECTIVES, ADVERB, NOUN PHRASE, VERB PHRASE etc. each and every grammar category has a priority in statement. So our approaches to extract out the Metadata extraction based on its priority in grammar. Priority in grammar component is as follows: noun, verb, adjective, adverb, noun phrase

Proposed Idea

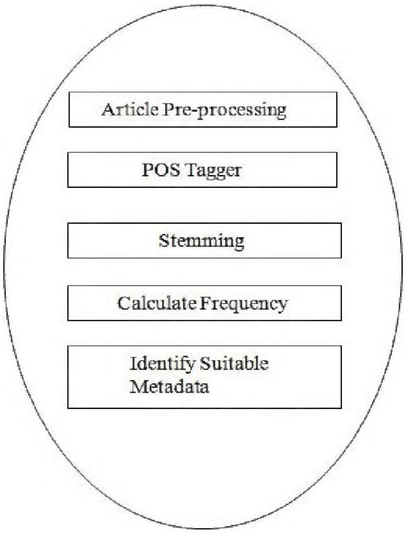

Figure 1: Proposed System Architecture

In figure-1 we give proposed system architecture. In this architecture we does not stick steps in any order.

Article Pre-Processing

Article pre-processing which remove irrelevant contents (i.e. tags, header-footer details etc.) from documents.

POS Taggers

A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some languages and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc.

Stemming

In most cases, morphological variants of words have similar semantic interpretations can be considered as equivalent for the purpose of IR applications. For this reason, a number of so-called stemming Algorithms, or stemmers, have been developed, which attempt to reduce a word to its stem or root form.

Calculate Frequency

Here each termed frequency is calculated i.e. how many occurrence of each term in document.

Identify Suitable Metadata

Now metadata is extracted from word set based on their frequency, grammar and their positions.

Experiments & Results

In this study we take a corpus with 100 documents. Documents contain the news article about various categories. Here we first extract the metadata manually from each & every documents. Then apply our idea to corpus. We measure our result from following parameter.

Precision = No of terms identified correctly by the system / Top N terms out of total terms generated by the system.

Recall = Number of keyterms identified correctly by the system / Number of keyterms identified by the authors.

F-measure=F= 2* ((precision* recall)/ (precision+ recall))

Table 1: Evaluation Results

Terms	Precision	Recall	F-measure
10	0.43	0.36	0.40
20	0.42	0.63	0.51
30	0.32	0.72	0.49

CONCLUSIONS & FUTURE WORKS

This method based on grammar component Our Aim to use this algorithm to identifying metadata in bigram, trigram tetra gram. This metadata helps us to generate summary of documents.

REFERENCES

1. Christopher D. Manning, Prabhakar, Raghavan, Hinrich Schtze an Introduction to Information Retrieval book.
2. H. P. Luhn. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. IBM Journal of Research and Development, 1957, 1(4): 309-317.

3. G. Salton, C. S. Yang, C. T. Yu. A Theory of Term Importance in Automatic Text Analysis, Journal of the C. Zhang et al American society for Information Science, 1975, 26(1): 33-44.
4. Y. Matsuo, M. Ishizuka. Keyword Extraction from a Single Document Using Word Co-ocuurrence Statistical Information International Journal on Artificial Intelligence Tools, 2004, 13(1): 157-169.
5. Yan Gao Jin Liu, Peixun Ma The HOT keyphrase Extraction based on TF*PDF, IEEE conference, 2011.
6. Canhui Wang, Min Zhang, Liyun Ru, Shaoping Ma An Automatic Online News Topic Keyphrase Extraction System,IEEE conference, 2006.
7. Nor Adnan Yahaya, Rosiza Buang Automated Metadata Extraction from web sources, IEEE conference, 2006.
8. Somchai Chatvienchai Automatic metadata extraction classi_cation of spreadsheet Documents based on layout similarities, IEEE conference, 2005.
9. Dr. Jyoti Pareek, Sonal Jain KeyPhrase Extraction tool (KET) for semantic metadata annotation of Learning Materials, IEEE conference, 2009.
10. Wan Malini Wan Isa, Jamaliah Abdul Hamid, Hamidah Ibrahim, Rusli Abdullah, Mohd. Hasan Selamat, Muhamad Tau_k Abdullah and Nurul Amelina Nasharuddin Metadata Extraction with Cue Model.
11. Zhixin Guo, Hai Jin A Rule-based Framework of Metadata Extraction from Scienti_c Papers, IEEE conference.
12. Ernesto Giralt Hernndez, Joan Marc Piulachs Application of the Dublin Core format for automatic metadata generation and extraction, DC-2005: Proc. International Conference. On Dublin Core and Metadata Applications.
13. Canhui Wang, Min Zhang, Liyun Ru, Shaoping Ma An Automatic Online News Topic Keyphrase Extraction System, IEEE conference.
14. Srinivas Vadrevu, Saravanakumar Nagarajan, Fatih Gelgi, Hasan Davulcu Automated Metadata and Instance Extraction from News Web Sites, IEEE conference.